Perception, 1995, volume 24, pages 1227–1232

Guest editorial

Vision is getting easier every day

You may have noticed, as I have, that the complexity of the visual system oscillates over time. For quite a long time, centuries in fact, vision could call on rich internal representations (Al-Haytham c. 1030, English translation Sabra 1989; Helmholtz 1867; Koffka 1935). The visual system was seen as a substantial piece of work, but this view ended in 1950 when Gibson claimed that vision did not require any internal representations. The richness was in the world, he said, and the visual system simply resonated to it. A little later, the new field of AI and cognitive science proposed a similarly scaled-down visual system. Vision was merely a front end. It appeared as a very small box in most diagrams that I dutifully copied down during my graduate school courses. The box delivered straightforward descriptions of the scene to a complex information processor which did the real work. Clearly the brain had to be as complex as the world it dealt with, but the visual system was a piece of cake.

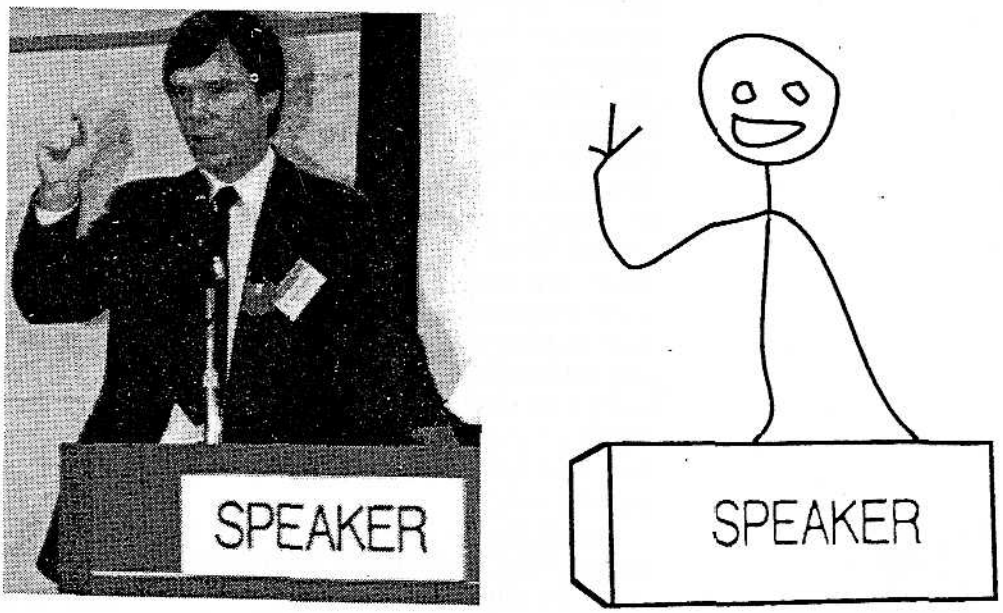
In fact, vision was so simple, goes a story we have all heard many times, a graduate student was delegated to program a visual system as a summer project at MIT (Roberts 1965). With this birth of computer vision, the complexity of vision gradually returned to impressive dimensions. Computational approaches uncovered a range of new operations of great power—object-centered representations, regularization, wavelets, relaxation labeling, constraint satisfaction—all of which were useful in analyzing surfaces and objects and in building a three-dimensional model of the scene. However, this work also revealed that these complex tools were not good enough. The overall impression was that vision was not only hard but even too hard for the best computational methods.

We are now on the cusp of a return swing and claims of simplicity are breaking out all over. It is once more possible to read the word 'template' in articles about object recognition (albeit, more sophisticated templates). Several papers propose that recognition may depend on 2-D views rather than general 3-D models of objects (Rock et al 1981; Tarr and Pinker 1990; Cavanagh 1991; Bülthoff and Edelman 1992; Logothetis et al 1995). Others propose that our interpretations of 3-D surface structure may be learned from patterns of associations, not computed from image structure (Nakayama and Shimojo 1992). Our depth judgements are sometimes so poor that our representation of depth cannot be metric and may be only a little better than ordinal (Todd and Reichel 1989). As the extreme in this trend, O'Regan (1992) claims that there is only the most minimal representation of the visual world, that the world itself serves as its own representation—after all, it is always there and if we want to know something about it, we just go look. The idea that the world is the external memory for vision is based on the point that—unlike, say, speech sounds—the world does not go away. The same could be said of pain. We do not need an internal memory for where our body hurts. Say that some weasels were gripping onto our body at a few different spots. We would not have to ask ourselves "now where are those weasels?"

I would like to take a brief look at three aspects of this minification of the visual system. First, if vision is so simple, why don't we know how it works? Second, if it is so simple, why does the world look so convincing? Last, is this simplicity the reason artists can get away with so much?

If vision is so simple, why don't we know how vision works? You might want to argue that if decades of research by hundreds of highly trained professionals have not revealed the answer, vision must be complex. But many simple things remain perplexingly out of reach. Romans had steam power but didn't use it. The Incas had wheeled toys but no wheeled vehicles. We put our garbage into soggy paper grocery bags until the mid 1960s even though plastic bags had become available many years earlier. The headlights on Italian cars turn off when the key is removed from the ignition, but not so on American cars—the battery dies instead. Post-Its could easily have been a Victorian-era invention. Clearly, simplicity is no guarantee of discovery by humans. But is vision as a whole simple, or just the parts that the proponents of the simple vision have addressed—3-D structure, stored representations of objects, memory of the world?

It would seem that there are too many parts to the visual cortex, just too many neurons, for the whole system to be considered simple. If we cannot describe the whole edifice as simple, is there some core theory of vision which might be considered simple? Our best analogy for a theory of vision at the moment is, say, a shelf full of operating manuals for a mainframe computer, or as others put it, a Swiss Army knife (Cosmides et al 1992) or a bag of tricks (Ramachandran 1985). But these are not theories. If we were discussing living organisms, we could similarly compile a user's manual of proteins that make livers function, enzymes that aid digestion, and lipids that make membranes. We would not call this a theory of living organisms; instead, we would look to the genetic code of DNA as the information storehouse for all these functions. Similarly, for a theory of vision, we should look to a theory of the representation of visual information: the image formats that allow communication between the visual cortices and which allow compact descriptions of visual events to be broadcast from centers of vision to centers of planning and storage. This will not be a theory of the computational goals and implementation of an assortment of visual algorithms and heuristics—that is too much like a user's manual or more



What we think we see What we really see

Figure 1.

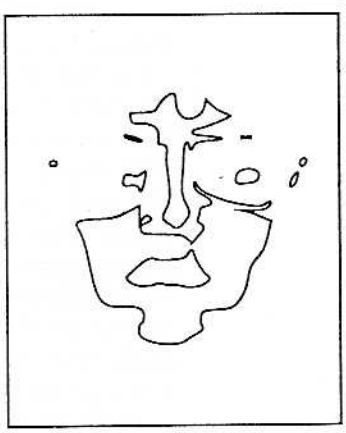
precisely a recipe for writing a user's manual. Having said that, I realize that it says little about whether or not this code will ultimately be a simple code. So for the moment the move to a simpler visual system can only claim to apply to its parts and not its whole. Nevertheless, the simplicity of these parts raises challenges to our intuitions about vision.

Our world appears to us as a complex three-dimensional volume of surfaces, stuff, and light. How could such a finely wrought sensory experience be a ruse? Moreover, on a practical level, things are where we expect them to be when we reach for them; they feel like they look. The claim of simple vision is that the real representation of the visual world is coarse, crude, and sparse, but that over this flimsy structure is draped a beautifully detailed 2-D texture. Remember that the 2-D array that falls on the retina is hardly changed at all by subsequent analyses—true, brightness and colours are normalized, but the 2-D position of any particular image feature is rarely affected. The visual system works mightily to invent a 3-D structure and identify objects to explain the 2-D patterns but even if this were very crude, the 2-D texture map remains detailed and true to the world. Add to this the ability to fill in details from memory, to interrogate extraretinal signals like accommodation and vergence and the illusion becomes compelling.

Is this underlying simplicity the reason that artists can get away with so much? Is simplicity in art evidence for simplicity in the processes which decode art? Art which only reproduces the visual world tells about the world but little or nothing about the brain: Sculptures *are* 3-D, holograms appear so, photographs and photorealism resemble the scenes they capture. On the other hand, art which captures the 3-D structure of the world without merely recreating or copying it may offer a revealing glimpse of the inner code of vision. Let's call these techniques pseudorealism because it is pleasingly oxymoronic. Obvious examples are the lines used in line drawings, the flatness of paintings and photographs, and the stark simplicity of two-tone depictions (see figure 2). These features are so commonplace that we seldom question the reason why they work.

A line drawing of a house, or a car, or a lion can look very convincing but remember that there are no lines in the real world corresponding to the lines used in the drawings. We can never have learned from experience that these lines stand for depth discontinuities—folds or occlusions in the surfaces depicted. In the real world, these depth discontinuities are revealed by changes in brightness, texture, color, or

Two-tone image



Contours of same image

Figure 2.

whatever but these are edges with one value extending on one side and a different value on the other. This is not a line. Why should lines work at all? We can reject the simple notion that line drawings are a learned convention, passed on through our culture. This point has been a contentious one (Kennedy 1975; see Deregowski 1989, and following comments) but more recent evidence points to the conclusion that line drawings are universally interpreted in the same way—infants, stone-age tribesmen and even monkeys appear to be capable of interpreting line drawings as we do. Nor is it the case that the lines in line drawings simply trace the brightness discontinuities in the image, because this type of representation is rendered meaningless by the inclusion of cast shadow and pigment contours (see figure 2, right panel). By a quirk of design or an economy of encoding, lines may be directly activating the internal code for object structure, but only object contours can be present in the drawing for this shortcut to work. It is as if we have stripped away the facade and can experience the simplicity of the structure which underlies the vision of rich natural scenes without the necessity of draping the structure with its complex 2-D texture. If we accept this view, then our inner representations may be as skimpy as a sketch, and on this count the visual system seems simple.

If line drawings lay bare our simple internal codes, what do we make of the flatness of paintings and photographs? Specifically, flat paintings provide consistent, apparently 3-D interpretations from a wide range of view points. This is not only convenient for the artist, but also prime evidence that our impressions of a 3-D world are not supported by true 3-D internal representations. Marvin Minsky claimed that if we had true 3-D vision, Rubik's cube would have been a boringly trivial game. If we had real 3-D vision, the object depicted in the flat picture would have to distort grotesquely in 3-D space as we moved about the picture. After all, consider what happens when we move around an inverted mask while perceiving it in reversed depth. The object appears to rotate with us, actually at twice our speed, and we are startled and amused. With a picture, the 3-D object we are seeing also has to rotate with us, but at the same speed we are moving, in order to present the same 'face' toward us as we move. It also must compress and expand horizontally as we move. These distortions should therefore be almost as dramatic as those seen for the inverted mask. To the contrary, however, objects in pictures seem reassuringly the same as we change our vantage point (with some interesting exceptions, see Gregory 1994). We don't experience the distortions—say proponents of simple vision—because we don't have a real 3-D representation of the object. It has some qualities of three dimensions but is far from metrically appropriate. The effectiveness of flat images is of course a boon to artists, who do not have to worry about special vantage points, and to film makers, who can have theatres with more than one seat in them.

Pictures can also be very sparse and an extreme example is the two-tone image style which emerged in the world of graphic arts at the end of the last century. In these images, there is a minimum of information and yet the object has a compelling 3-D structure (Hayes 1988). Here are some of the interesting properties of these images. First, there is no counterpart to these images in the real world—there are no circumstances which give rise to two-valued images for real 3-D objects. A dark object silhouetted against a bright background comes close, but silhouettes lack internal detail whereas two-tone images like that in figure 2 have complex internal detail. Since monkeys can also recognize two-tone images such as these [or at least their face-sensitive cells do, after one exposure to the grey-scale version of the same image (Rolls et al 1993)], we cannot appeal to learned conventions of our culture. These images cannot be interpreted on the basis of the local structure of their outlines—the contours alone often appear as meaningless scribbles. The parts and features of the objects in the images cannot be determined from intersections,

junctions, or deep concavities (try it). These images are impossible to interpret with the aid of any part-based model (eg, geons, generalized cylinders, or superquadrics), structural encoding (eg, medial axes, cores, or codons), distinctive features (colours or textures), or depth recovery process (texture gradients, disparity, or pictorial cues). What can possibly be left if no parts can be found, if no depth relations can be determined, and if it is unknowable whether any contour belongs to an object or to a cast shadow? Quite likely, the only remaining process is the simplest of all pattern operations: viewpoint-specific recognition. This is, undeniably, a modern code word for 2-D templates. Parts of the pattern match a particular 2-D view of, in this case, a face—perhaps a generic face. Once a match is found, other parts of the pattern can be interpreted in that context. As would be expected for such a top-down process, it only works for familiar objects seen from familiar viewpoints. A two-tone image of an unfamiliar structure such as, say, a mountain range or lump of clay, does not give rise to any impressions of 3-D structure. Interestingly, this primitive, view-specific form of object recognition is able to operate on its own when presented with two-tone images of familiar objects. It is not countermanded by the impossibility of a two-tone image in the real world. This technique is a relatively recent discovery in the 40 000 year history of art. Like many aspects of art, it informs us about the brain within us as much as about the world around us. In particular, two-tone images, along with line drawings and the viewpoint independence of paintings, all point to an underlying simplicity in visual processes.

The trend towards a simpler visual system may be short-lived but for the moment we can enjoy the 'Just So' stories that it offers. The claims for the simplicity of vision, given the immense task it faces reminds me of a parable. In this parable, or one like it, vision is like a Venus's-flytrap trying to eat a frog. It bites and bites but it only has little plant teeth. It never gets the whole frog but other stuff happens and that is vision.

Patrick Cavanagh

References

Bülthoff H H, Edelman S, 1992 "Psychological support for a two-dimensional view interpolation theory of object recognition" *Proceedings of the National Academy of Sciences of the USA* **89** 60–64

Cavanagh P, 1991 "What's up in top-down processing?", in *Representations of Vision: Trends and Tacit Assumptions in Vision Research* Ed. A Gorea (Cambridge, UK: Cambridge University Press) pp 295–304

Cosmides L, Tooby J, Barkow J, 1992 "Introduction: Evolutionary psychology and conceptual integration", in *The Adapted Mind* Ed. J Barkow et al (Oxford: Oxford University Press) pp 3–15

Deregowski J B, 1989 "Real space and represented space: cross-cultural perspectives" *Behavioral and Brain Sciences* **12** 51–119

Gibson J J, 1950 *The Perception of the Visual World* (Boston: Houghton Mifflin)

Gregory R, 1994 "Experiments for a desert island" *Perception* **23** 1389–1394

Hayes A, 1988 "Identification of two-tone images: some implications for high- and low-spatial-frequency processes in human vision" *Perception* **17** 429–436

Helmholtz H von, 1867 *Handbuch der physiologischen Optik* 1st edition (Leipzig: Voss)

Kennedy J M, 1975 "Drawings were discovered, not invented" *New Scientist* **67** 523–527

Koffka K, 1935 *Principles of Gestalt Psychology* (New York: Harcourt, Brace)

Logothetis N, Pauls J, Poggio T, 1995 "Shape representation in the inferior temporal cortex of monkeys" *Current Biology* **5** 552–563

Nakayama K, Shimojo S, 1992 "Experiencing and perceiving visual surfaces" *Science* **257** 1357–1363

O'Regan J K, 1992 "Solving the 'Real' mysteries of visual perception: the world as an outside memory" *Canadian Journal of Psychology* **46** 461–488

Ramachandran V S, 1985 "The neurobiology of perception" *Perception* **14** 97–105

Roberts L G, 1965 "Machine perception of three-dimensional solids", in *Optical and Electro-optical Information Processing* Eds J T Tippett et al (Cambridge, MA: MIT Press) pp 159–197

Rock I, DiVita J, Barbeito R, 1981 "The effect on form perception of change of orientation in the third dimension" *Journal of Experimental Psychology: Human Perception and Performance* **7** 719–732

Rolls E T, Tovee M J, Ramachandran V S, 1993 "Visual learning reflected in the response of neurons in the temporal visual cortex of the macaque" *Society for Neurosciences Abstracts* **19** 28

Sabra A I, 1989 *The Optics of Ibn Al-Haytham* English translation (London: Warburg Institute, University of London)

Tarr M, Pinker S, 1990 "When does human object recognition use a viewer-centered reference frame?" *Psychological Science* **1** 253–256

Todd J T, Reichel F D, 1989 "Ordinal structure in the visual perception and cognition of smoothly curved surfaces" *Psychological Review* **96** 643–657

p

© 1995 a Pion publication printed in Great Britain